Big Data - Mise en oeuvre de traitements avec Spark

Mettre en oeuvre Spark pour optimiser des calculs

Présentiel

3 jours (21 h) | Prix inter : 2.590,00 € HT | Réf.: BD515

L'essor du Big Data a considérablement fait évoluer l'écosystème Hadoop, à l'origine principalement constitué de HDFS et MapReduce. Parmi les nouveaux outils qui gravitent autour d'Hadoop, Apache Spark, framework dédié au traitement et à l'analyse de données massives, a particulièrement attiré l'attention à tel point que quelques mois après sa mise à disposition sur le marché, les fournisseurs de solutions Hadoop l'ont intégré à leurs distributions. S'il rencontre un franc succès, c'est bien que Spark se pose en alternative crédible à MapReduce dont la mise en oeuvre est parfois lourde. En effet, contrairement à MapReduce, Spark propose un framework complet et unifié pour répondre aux besoins de traitements de données hétérogènes tout en permettant aux applications Hadoop d'être exécutées beaucoup plus rapidement.

A qui s'adresse cette formation ?

- Chefs de projet
- Data Scientists
- Développeurs

- Connaissance de Java ou Python
- Avoir suivi le séminaire "Hadoop - Présentation de l'écosystème" (SEM35) ou avoir des bases Hadoop
- Notions de calculs statistiques

Programme

Pendant

1 - Introduction

- Présentation de Spark
- Origine du projet
- Apports et principes de fonctionnement
- Langages supportés
- Mise en oeuvre sur une architecture distribuée
- Architecture : clusterManager, driver, worker, ...

2 - Premiers pas

- Utilisation du Shell Spark avec Scala ou Python
- Modes de fonctionnement
- Interprété, compilé
- Utilisation des outils de construction
- Gestion des versions de bibliothèques
- Mise en pratique en Java, Scala et Python
- Notion de contexte Spark
- Extension aux sessions Spark

3 - Règles de développement

- Mise en pratique en Java, Scala et Python
- Notion de contexte Spark
- Différentes méthodes de création des RDD : depuis un fichier texte, un stockage externe
- Manipulations sur les RDD (Resilient Distributed Dataset)
- Fonctions, gestion de la persistance

4 - Cluster

- Différents cluster managers : Spark interne, avec Mesos, avec Yarn, avec Amazon EC2
- Architecture : SparkContext, SparkSession, Cluster Manager, Executor sur chaque noeud
- Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job
- Mise en oeuvre avec Spark et Amazon EC2
- Soumission de jobs, supervision depuis l'interface web

5 - Traitements

- Lecture/écriture de données : texte, JSon, Parquet, HDFS, fichiers séquentiels
- Jointures
- Filtrage de données, enrichissement
- Calculs distribués de base
- Introduction aux traitements de données avec map/reduce

6 - Support Cassandra

- Description rapide de l'architecture Cassandra
- Mise en oeuvre depuis Spark
- Exécution de travaux Spark s'appuyant sur une grappe Cassandra

7 - DataFrames

- Spark et SQL
- Objectifs : traitement de données structurées
- L'API Dataset et DataFrames
- Optimisation des requêtes
- Mise en oeuvre des Dataframes et DataSet
- Comptabilité Hive
- Travaux pratiques : extraction, modification de données dans une base distribuée
- Collections de données distribuées
- Exemples

8 - Streaming

- Objectifs , principe de fonctionnement : stream processing
- Source de données : HDFS, Flume, Kafka, ...
- Notion de Streaming
- Contexte, DStreams, démonstrations
- Travaux pratiques : traitement de flux DStreams en Scala
- Watermarking
- Gestion des micro-batches
- Travaux pratiques : mise en oeuvre d'une chaîne de gestion de données en flux tendu (IoT, Kafka, SparkStreaming, Spark)
- Analyse des données au fil de l'eau

9 - Intégration Hadoop

- Rappels sur l'écosystème Hadoop de base : HDFS/Yarn
- Création et exploitation d'un cluster Spark/YARN
- Intégration de données sqoop, kafka, flume vers une architecture Hadoop et traitements par Spark
- Intégration de données AWS S3

10 - Machine Learning

- Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistance, statistiques
- Mise en oeuvre avec les DataFrames

11 - Spark GraphX

- Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes
- Travaux pratiques : exemples d'opérations sur les graphes

Les objectifs de la formation

- Pouvoir comprendre le fonctionnement de Spark et son utilisation dans un environnement Hadoop
- Savoir intégrer Spark dans un environnement Hadoop
- Comprendre comment traiter des données Cassandra, HBase, Kafka, Flume, Sqoop et S3

Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques.
Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Les points forts de la formation

- Une formation qui accorde une large place à la pratique : de nombreux exercices seront réalisés tout au long de la formation.
- Les retours d'expérience et conseils de consultants experts du domaine.
- 83% des participants à cette formation se sont déclarés satisfaits ou très satisfaits au cours des 12 derniers mois.

Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques.
Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Dates et villes 2027 - Référence BD515

Dernières places disponibles Session garantie

Paris

du 15 mars au 17 mars	du 20 sept. au 22 sept.
du 27 mai au 29 mai	du 6 déc. au 8 déc.